Fall 2021

Group 6

Romnick Galang

Shaniquo McKenzie

Mustafa Butt

Fawad Sayed Rahim

COMP309 – Assignment 2

Fall 2021



Predictive Machine Learning Model On Bicycle Theft Data

**COMP 309 Data Warehouse & Predictive Modeling**

**Predictive Machine Learning Model**

Logistic Regression Model

Group 6

Romnick Galang

Shaniquo McKenzie

Mustafa Butt

Fawad Sayed Rahim

Contents

[**1.0 Executive summary 3**](#_Toc90676946)

[**1.1 Problem 3**](#_Toc90676947)

[**1.2 Solution 3**](#_Toc90676948)

[**1.3 Key Findings 4**](#_Toc90676949)

[**2.0 Overview 5**](#_Toc90676950)

[**3.0 Data exploration and findings 5**](#_Toc90676951)

[**3.1 Dataset field descriptions 5**](#_Toc90676952)

[**3.2 Graphs and Visualizations 7**](#_Toc90676953)

[**3.3 Tools & Languages 9**](#_Toc90676954)

[**3.4 Libraries and Frameworks 10**](#_Toc90676955)

[**4.0 Feature selection 11**](#_Toc90676956)

[**4.1 Feature Selection- Scikit learn RFE package 11**](#_Toc90676957)

[**4.2 Feature Selection - Business logic 12**](#_Toc90676958)

[**5.0 Data modeling 14**](#_Toc90676959)

[**5.1 Data Cleaning 14**](#_Toc90676960)

[**5.2 Standardize the data 15**](#_Toc90676961)

[**6.0 Model building 15**](#_Toc90676962)

[**6.1 Model Selection 15**](#_Toc90676963)

[**6.2 Train/Test data 16**](#_Toc90676964)

[**6.3 Confusion Matrices 16**](#_Toc90676965)

# 1.0 Executive summary

## 1.1 Problem

Approximately 4000 bicycles have been stolen every year since 2015 in Toronto. The bicycle theft rate has increased every year since 2014. With the advent of electric and motorized bikes in the era of environmentalism, bicycles are increasing in popularity and numbers. Increases in both thefts and numbers of bikes on the streets poses a potentially troubling trend. What happens to those bikes once they are stolen? Are the bikes ever recovered? If so, how would a victim know the odds of recovering their bike? What factors determine whether a stolen bike is likely to be recovered? Like with all questions, the answer begins with information. But where do we look for information that will help predict whether a bike will be recovered? The answer is raw data.

## 1.2 Solution

Our team had the chance to work with a Toronto Police Service’s bike theft dataset that recorded bicycle theft for the past seven years. The bike theft data was carefully evaluated, analyzed and processed before being used to develop a model that should predict whether a stolen bike will be recovered, based on key attributes of the bike theft. Looking at recent historical data that includes dates of occurrences and reporting, bike features and properties, and location details, we built our model to predict whether a stolen bike is likely to be recovered based on the combination of these attributes. Apart from using trends in the dataset to develop the model, the bike theft data was also used to both train and test the model, allowing us to find the ideal combination of attributes to most accurately predict the outcome.

## 1.3 Key Findings

During the stage of feature selection, while testing the developing model with different combinations of features, we gained insight into the importance of some features over others. Through a process of elimination, and trial and error, we further refined the model, along with our understanding. The key findings we were able to discover included:

* Majority of bike theft occurrences occur in the months between June-September which directly correlated with the months that had the highest frequency of recovery.
* Bikes of type MT were more likely to be recovered.
* Bikes stolen on a weekend are less likely to be recovered
* Bikes stolen, that worth around $3000
* Bikes stolen in the winter months are least likely to be recovered
* Premises types was most determinative in predicting and/or effecting recovery status

At the conclusion of the project, the team found out that certain factors contributed more heavily to determining whether a stolen bike will be recovered.. The team was able to develop a logistic regression model that can potentially provide valuable insight to people.

# 2.0 Overview

The team built a logistic regression model that is based on the bicycle theft dataset from the Toronto police department. The model was written in python using pandas as the main library. The dataset was cleaned and transformed such that missing values and imbalances were managed. Features that would be included in the model were selected through the use of the Scikit learn library. The accuracy scores were assessed throughout the project and methods were performed to increase its accuracy.

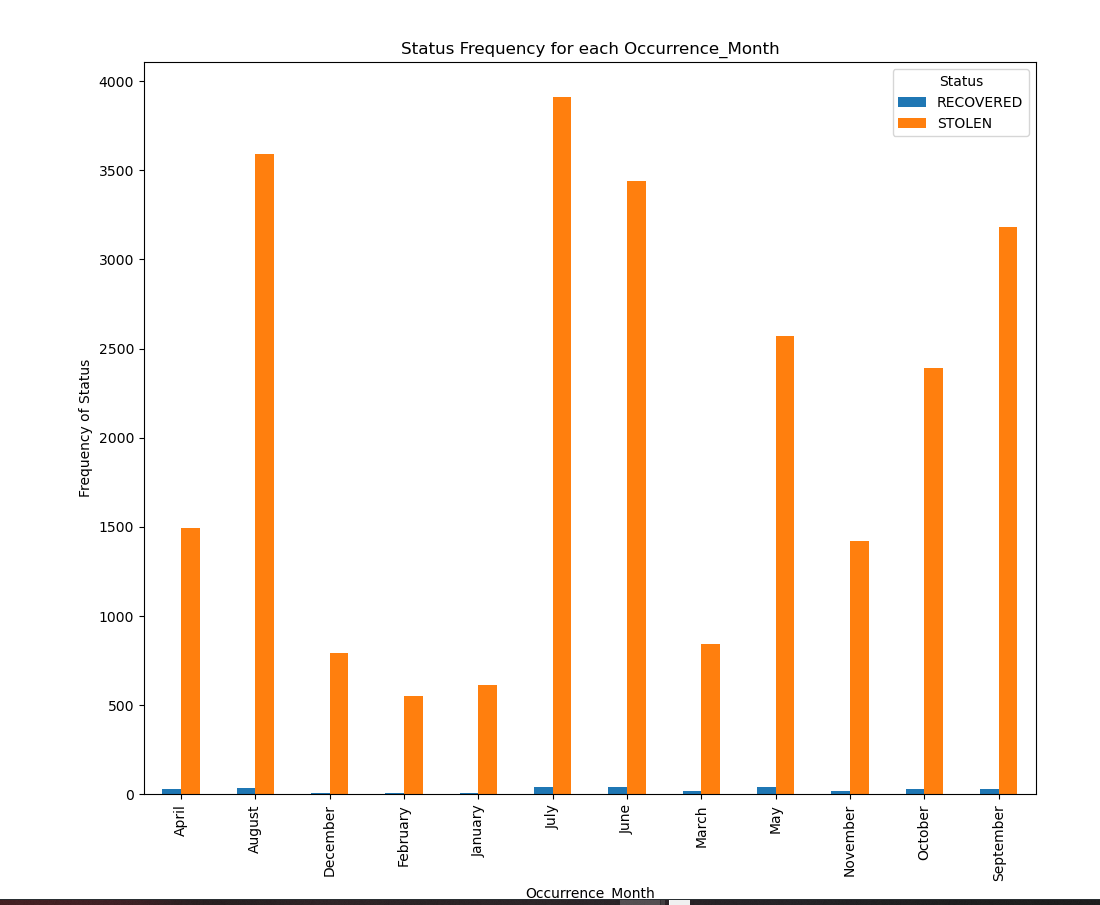
The model was serialized using Pickle for it to be stored in a local disk and be used to perform a prediction later. The model was deployed to an API using the Flask framework. Subsequently, Postman was used to test the model using the post method sending json objects to the model.

# 3.0 Data exploration and findings

## 3.1 Dataset field descriptions

|  |  |  |
| --- | --- | --- |
| **Field** | **Field Name** | **Description** |
| 1 | OBJECTID | Unique Identifier |
| 2 | event\_unique\_id | Occurrence Number |
| 3 | Primary\_Offence | Primary Offence Type |
| 4 | Occurrence\_Date | Date Theft Occurred |
| 5 | Occurrence\_Year | Year Theft Occurred |
| 6 | Occurrence\_Month | Month Theft Occurred |
| 7 | Occurrence\_DayOfWeek | Day of Week Theft Occurred |
| 8 | Occurrence\_DayOfMonth | Day of Month Theft Occurred |
| 9 | Occurrence\_DayOfYear | Day of Year Theft Occurred |
| 10 | Occurrence\_Hour | Hour Theft Occurred |
| 11 | Report\_Date | Date Theft was Reported |
| 12 | Report\_Year | Year Theft was Reported |
| 13 | Report\_Month | Month Theft was Reported |
| 14 | Report\_DayofWeek | Day of Week Theft was Reported |
| 15 | Report\_DayOfMonth | Day of Month Theft was Reported |
| 16 | Report\_DayOfYear | Day of Year Theft was Reported |
| 17 | Report\_Hour | Hour Theft was Reported |
| 18 | Division | Police Division where Theft Occurred |
| 19 | City | City where Theft Occurred |
| 20 | Hood\_ID | Identifier of Neighbourhood where Theft Occurred |
| 21 | Neighbourhood | Name of Neighbourhood where Theft Occurred |
| 22 | Location\_Type | Location Type of Occurrence |
| 23 | Premises\_Type | Premises Type of Occurrence |
| 24 | Bike\_Make | Make of Bicycle |
| 25 | Bike\_Model | Model of Bicycle |
| 26 | Bike\_Type | Type of Bicycle |
| 27 | Bike\_Speed | Speed of Bicycle |
| 28 | Bike\_Colour | Colour of Bicycle |
| 29 | Cost\_of\_Bike | Cost of Bicycle |
| 30 | Status | Status of Bicycle |
| 31 | Longitude | Longitude Coordinates (Offset to nearest intersection) |
| 32 | Latitude | Latitude Coordinates (Offset to nearest intersection) |

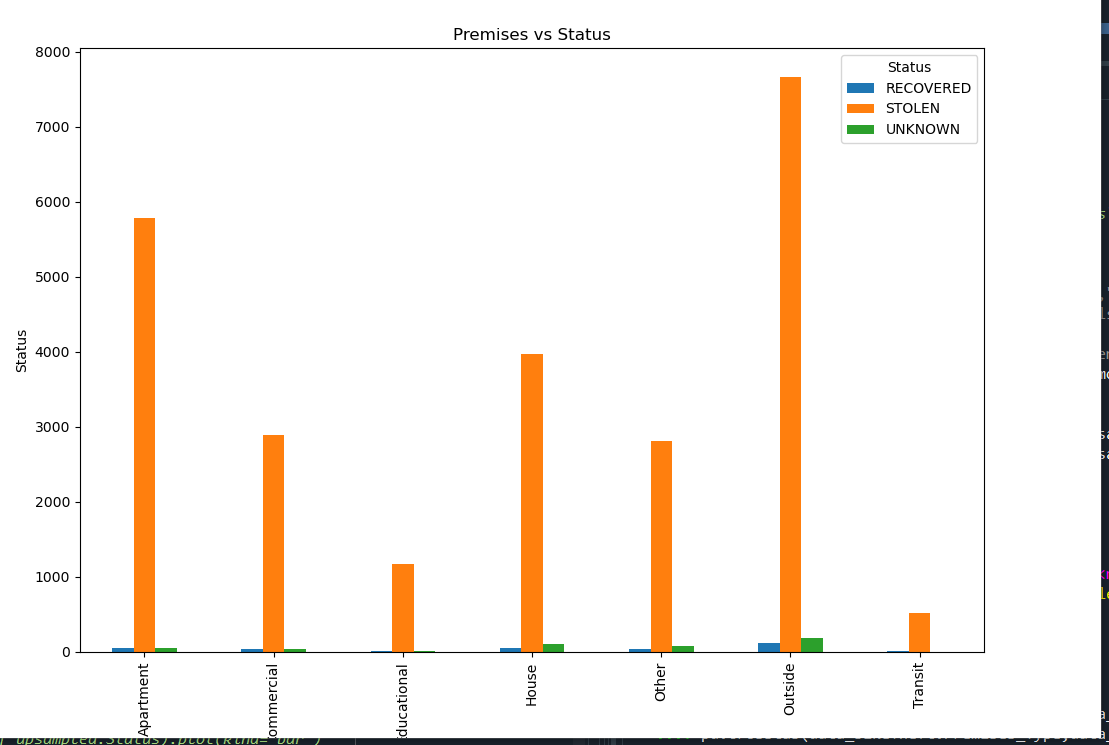
## 3.2 Graphs and Visualizations

In the histogram below bicycle theft crimes mostly occurred in the months of May, June, July, August and September. Obviously these months the weather is not that cold therefore people are using their bike more often

In the line graph below, the dataset tells us that the most occurrences were on the weekdays. Logically, these are the school days and office work days which explains the higher rates of bicycle theft during these days.

Chart, line chart

Description automatically generated

In the histogram below, it shows that most of the bicycles that were stolen were at the outside premises and the lowest rates were on the transit. Evidently, this means that bicycles that were outside were unattended and the ones in transit say otherwise.

## 3.3 Tools & Languages

**Spyder**- is a free and open source IDE which is written in python and for python. This tool was used to mainly explore and build the model.

**Postman**-is a graphical user interface that is mainly used to test API. This tool was used to test the API of the project using HTTP requests.

**Icon

Description automatically generatedPython** – is an interpreted high-level general-purpose programming language. It is a general-purpose, versatile, and powerful programming language.

## 3.4 Libraries and Frameworks

A picture containing text, clipart

Description automatically generated**Pandas-** this package was used to read the csv bicycle theft dataset and manipulate data on tables called DataFrames and Series. Pandas is a library built on NumPy. Data transformations and manipulations, and dummy variables were made possible with this library.

**Matplot**-this 2D plotting library was used to visualize the data with histogram, pairplot, scatter plot and boxplot. Plots included distribution, histogram, bar, line box and pair plots. With this library, we were able to nicely visualize data, for easier recognition of patterns and trends.

**NumPy**-this library is the core library for scientific computing. Core functionality relates to handling and manipulating arrays. This library was used with Pandas for processing and evaluating Series, columns and rows.

**Seaborn-** is a visualization library based on matplotlib but with greater display control. This library was used to create various plots including heatmaps, pairplots, scatterplots, boxplots and countplots.

**Sci-kit Learn**- is a library that enables to do a range of machine learning, preprocessing, cross-validation and visualization algorithms. This library was used to build the logistic regression model, which was used in predictive model building, standardization and in feature selection.

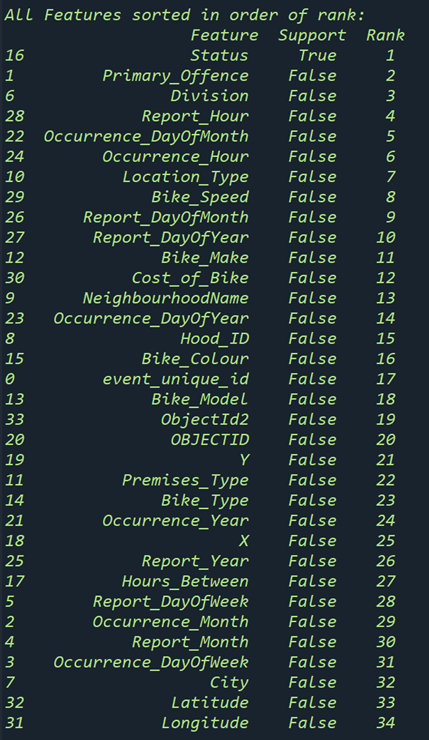
**Joblib**- was used to persist the predictive model into a file and store it in local drive as a pkl.

**Flask**- this framework was used to deploy the model which is built in server that runs the Flask web app.

# 4.0 Feature selection

## 4.1 Feature Selection- Scikit learn RFE package

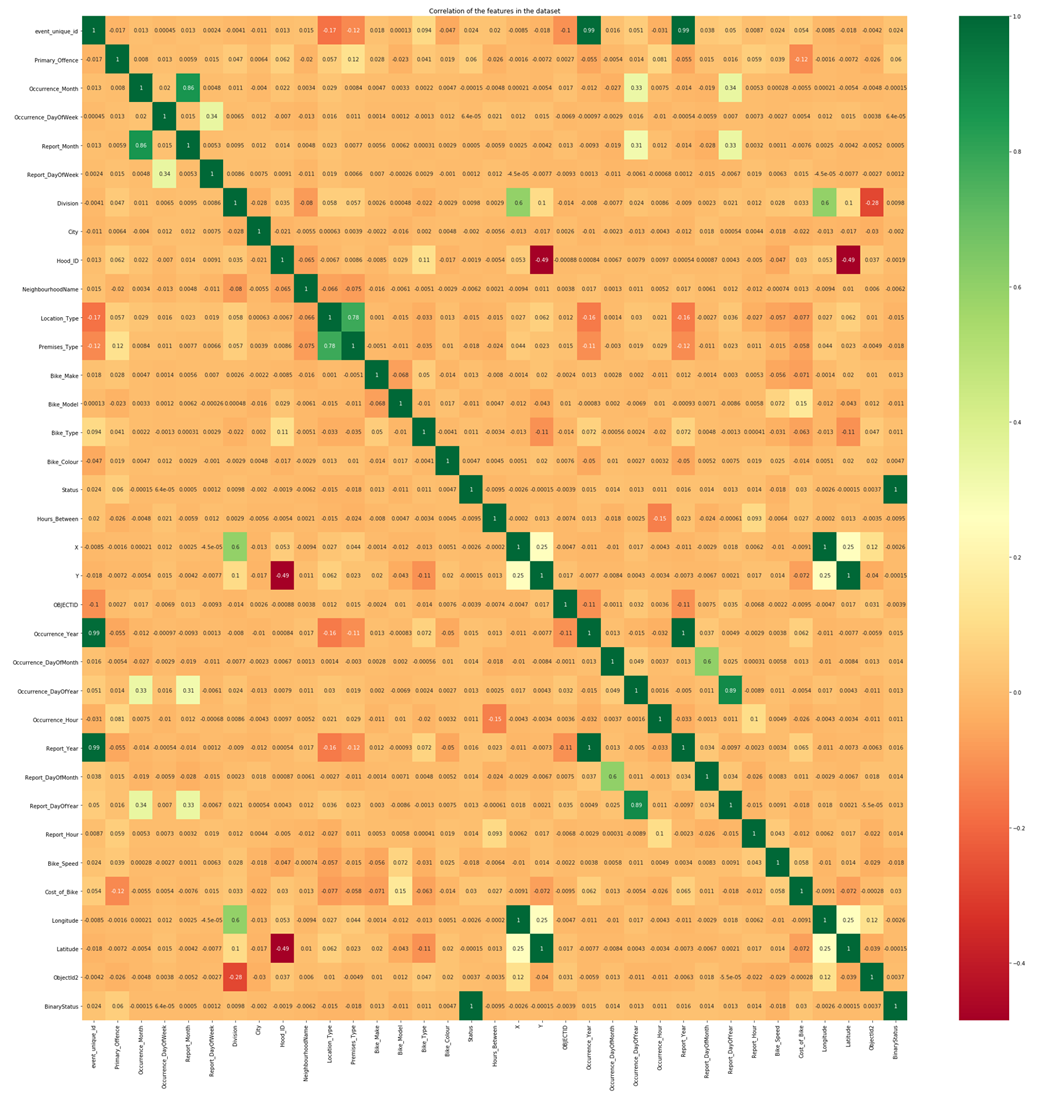
Feature selection in Python is the way in which we reduce the number of values used to determine whether a select feature or column is a good determination of the model or outcome. Basically it can tell whether or not a selected column is a good predictor of the target. There are multiple ways in which we can do feature selection for the data set which may include business logic, correlation or other statistics.

For our project, we will be exploring the data set called bicycle theft data where we will see which columns are the best predictor of the status of the bicycles. For our feature extraction method, we are using the logistic regression model and RFE packages to select the best features used to predict the status of the bikes. The packages, logistic regression and RFE, are both imported from the sklearn library in Python. Once the model is fitted and the desired number of features are selected, the model can be applied to the selected columns to predict whether or not the bicycle will have a status of stolen or recovered.

When using this method, we evaluated all the columns within the data set to see whether or not they are a good determination of the bicycle’s status. That way we can choose whichever columns are the best predictors or whether or not the bicycle will be stolen. When this method of feature extraction was applied to our data set, the rank for feature determination gave the order of importance for each column in determining whether or not they are good predictors of status. From this we can select the best columns based on the rank as well as the business logic.

## 4.2 Feature Selection - Business logic

Some business logic includes, the amount of time between occurrence and reporting the incident, the location (city, neighbourhood) where the incident took place, the month of the year, as well as the features of the bike (color, model). These are important features to consider as they could give some insights as all often bicycle theft occurred for each of these features.

 Image Showing Pearson correlation between columns in the dataset that can be used in feature selection before applying the supervise logistic regression model.

A close-up of a magnifying glass

Description automatically generatedGenerally, the columns with highest correlation are good predictors. It is also possible to compare coefficients to select the best predictor however, the dataset has to be normalized before regression can be performed then the absolute value of coefficients can be used to determine how often the column value was able to predict the target. Basically, we can evaluate the last row of the plot above where the green colors are the best predictors of the status while those closer to the red colors are not so good predictors of status based on the Pearson correlation. Also, we see that the Status column is a 100% match which is expected since it is a copy of the binary status but using numbers. The second-best predictor is primary offense with a 6% correlation to status.

Benefits of feature selection:

* Reduces Overfitting: Less redundant data means less opportunity to make decisions based on noise.
* Improves Accuracy: Less misleading data means modeling accuracy improves.
* Reduces Training Time: fewer data points reduce algorithm complexity and algorithms train faster.

# 5.0 Data modeling

Data modeling is the process of creating a visual representation of either a whole information system or parts of it to communicate connections between data points and structures. An important part of data modeling is cleaning the data, making assumptions and other data wrangling techniques. This is to ensure that the data being explored is in a standard form prior to making a predictive model.

## 5.1 Data Cleaning

For this data set, some of the data modeling that we have done includes feature selection as discussed earlier called non data transformation such as handling missing data, training the data on managing imbalance data.

## 5.2 Standardize the data

To ensure that the data is in a standard form, The data must first undergo some changes. For this particular dataset first, we start by removing the ‘unknown’ values from the dataset. We also grouped some columns in the dataset, for example, finding the times between report date and occurrence date, and also grouping the bikes within a certain price range for each cost of bike. These strategies again correlate with the business logic where we could evaluate whether having longer hours between occurrence and report would've had an impact on whether or not the bikes were recovered or what is the most common price range of the bikes being declared stolen.

When working with large datasets, sometimes assumptions have to be made. One assumption that we made for this data set is that if a value is missing then the most frequently occurring value is enough to replace the missing value. This data wrangling strategy modifies rows of data and is used to clean the dataset of any missing values that could cause discrepancies when using this data set to create the model.

# 6.0 Model building

## 6.1 Model Selection

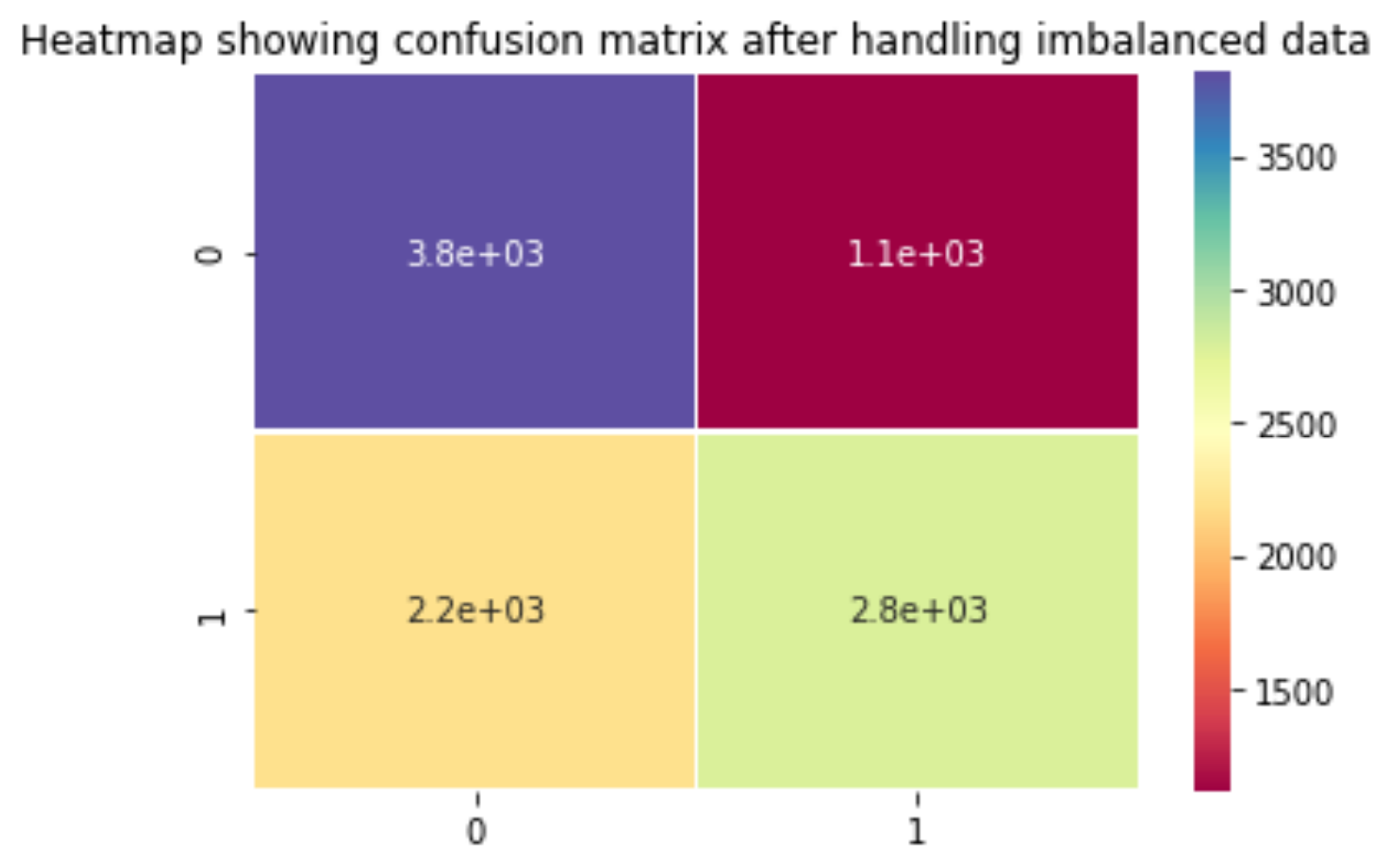
When building a model, it is ideal to select a predictive model that best fits the dataset for a more accurate determination of the target. For example, if our data set’ target is a categorical variable, then logistic regression can be used as well as decision trees. However, if the target is a continuous numeric variable, a linear regression model would be best to use. Also, we would need to train and test the data with whichever model is being chosen.

## 6.2 Train/Test data

For this project, we contort the dataset in such a way that it gives the best prediction score when it is being converted to the model. The bicycle theft dataset was Split into training and testing data. The data set is divided into 80/20%, where 80% is used for training and 20% of the data is used for testing. Some of the Python libraries included in the data wrangling pandas and sklearn. After the model is split by diving into the train and test data set, the logistic regression model is fit for the bicycle theft data set. When splitting the data set, the data is divided randomly however, only approximately 20% would be used for testing.

## 6.3 Confusion Matrices

After the dataset is fitted to the model, it is then used to predict the outcomes of target (status) in the test data. Once the model is fit and a prediction is complete, the results produces a confusion matrix which tells us the number of true positives, true negatives, false positives and false negatives that were generated for the target based on the selected columns.



References:

<https://globalnews.ca/news/7739568/bike-thefts-toronto-police/>

<https://data.torontopolice.on.ca/datasets/TorontoPS::bicycle-thefts/about>

<https://joblib.readthedocs.io/en/latest/generated/joblib.dump.html>

<https://palletsprojects.com/p/flask/>

<https://www.tutorialspoint.com/flask/flask_application.htm>

<https://www.datacamp.com/community/tutorials/pickle-python-tutorial#what>

[https://www.datacamp.com/community/tutorials/machine-learning-models-api-python](https://www.datacamp.com/community/tutorials/machine-learning-models-api-python?tap_a=5644-dce66f&tap_s=3575&utm_campaign=News&utm_medium=Community&utm_source=DataCamp.com)